Wrangle Report

By Hatem Kamal

13-10-2020

In this report I'll walk through the steps I took to complete this project. They are defined in three steps each step got its own time and effort.

1 – Gather

In this project we were required to get the data from 3 different sources to test 3 different methods of gathering data. First one was a provided csv file which I used the pandas function "read_csv()" to get, Next was downloading it from a link in tsv format which stands for tap separated values which I used "read_csv('', sep='\t')" to get, the third one was from Twitter API I couldn't get the API permission so I used an already downloaded file and extracted information using regular expression.

2 – Assess

After gathering the data comes the assessing where I get the problems in the data either visually or programmatically. Visually I used the csv file and txt file as the data wasn't very big. And programmatically I used the pandas dataframe function like "info()" and "value_counts()".

This process can be categorized into 2 issues Quality issue and Tidiness issue. After finding it I write it in a cell in the notebook.

Quality issue stands for the issues in the data content like missing data, missed typed data, wrong datatype etc.

The Tidiness issue is something about the data arrangement. For the data to be Tidy there're 3 condition:

1- Each column form a variable
2- Each row form a sample
3- Each value must have its own cell

3 – Cleaning

Here I clean the observed issues using regular expression again and some pandas functions like dropna(), drop(), merge(), melt(), etc.

Conclusion

Data wrangling is a really important step where it can form more than half of project process there're a lot of tools that can be used like pandas library in Python, or visually in excel to get the problem in the data after gathering it then clean it.